A Unified Modular Framework to Incorporate Structural Dependency in Spatial Omics Data

Jiayu Su[*1,2,3], Jean-Baptiste Reynier[1,4], Xi Fu[1,4], Guojie Zhong[2], Jiahao Jiang[5], Rydberg Supo Escalante[2], Yiping Wang[1,6], Benjamin Izar[1,6], David A Knowles[2,3,7], Raul Rabadan[*1,2,4]

[1]*Program for Mathematical Genomics, Columbia University, New York, NY, USA*
[2]*Department of Systems Biology, Columbia University, New York, NY, USA*
[3]*New York Genome Center, New York, NY, USA*
[4]*Department of Biomedical Informatics, Columbia University, New York, NY, USA*
[5]*Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK*
[6]*Department of Medicine, Division of Hematology/Oncology, and Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, NY, USA*
[7]*Department of Computer Science, Columbia University, New York, NY, USA*

[*] Correspondence: js5756@cumc.columbia.edu, rr2579@cumc.columbia.edu

Abstract
Spatial omics technologies, such as spatial transcriptomics, allow the identification of spatially organized biological processes, while presenting computational challenges for existing analysis approaches that ignore spatial dependencies. Here we introduce Smoother, a unified and modular framework that integrates positional information into non-spatial models via spatial priors and losses. In simulated and real datasets, we show that Smoother enables spatially aware data imputation, cell-type deconvolution, and dimensionality reduction with high accuracy.

Main
From subcellular arrangement to tissue compartmentalization, the spatial structure in an organism is highly organized at all scales. This harmonious architecture regulates a diverse variety of biological processes, including embryonic development, neuronal plasticity, and the tumor microenvironment. Recent advances in spatially resolved omics technologies provide unique opportunities to study the spatial patterns of gene and epigenetic activities and the dynamics of biological systems at the cellular and tissue level[1-5]. While existing non-spatial omics analysis methods can be applied to spatial data, the neglect of positional information makes them incapable of overcoming limitations such as low spatial resolution and structured technical noise[6], let alone inferring biologically meaningful spatial organization.

Here we present Smoother, a unified framework that integrates spatial information through the employment of spatial priors and losses. Smoother exploits the significant similarities of adjacent cells and locations consistently observed across tissues and technologies, with pairwise similarity decaying at different rates with distance (**Supplementary Fig. 1**). It is designed to be modular and thus compatible with a wide variety of applications, enabling the use of existing non-spatial single-cell based strategies in a spatially aware manner. By representing data as weighted graphs and Markov random fields[7, 8], Smoother explicitly characterizes the dependency structure to allow information exchange between neighboring locations and facilitates robust and scalable inference of cellular and cell-type activities. We demonstrate the versatility of Smoother by implementing and testing its performance on tasks including cell-type deconvolution and dimensionality reduction. Using simulated and real spatial omics data of different modalities, our benchmarks reveal the substantial advantages of explicitly modeling and incorporating spatial dependencies.

Smoother differs from existing models in that it treats data-specific dependencies as shared and reusable priors across downstream tasks to encourage local smoothness on any spatial variables of interest (**Fig. 1a, Methods and Supplementary Notes**). It first builds a spatial graph in which edges connecting physically adjacent locations are weighted and pruned using histological and transcriptomic similarities to reflect domain boundaries (**Extended Data Fig. 1**). The graph is then converted into a multivariate normal (MVN) prior with varying degrees of dependencies and an equivalent spatial loss that can be attached to non-spatial models (**Supplementary Fig. 2**). As a simple showcase of Smoother, we considered the problem of data imputation where the gene expression at unobserved locations is inferred from a hidden Markov random field with the MVN prior[9] (**Methods**). By penalizing local variations of expression over space, Smoother mitigated dropout effects and recovered spatial patterns of cortical layer markers in a human dorsolateral prefrontal cortex (DLPFC) dataset[10], which highlights the power of spatial context

(**Supplementary Fig. 3**). The same scheme can be applied to enhance the spatial resolution of the data to an arbitrary scale (**Supplementary Fig. 4**).

A common challenge for barcode-based spatial omics technologies that capture a mixture of cells at each location is to disentangle cell-type composition across space. Despite the many deconvolution methods[11-16] developed for spatial transcriptomics (ST) data, few recognize spatial dependency explicitly. Even for these methods, for instance CARD[17] and BayesTME[18], the spatial covariance is hard-coded and not transferable to other models. To fill this gap, we extended four non-spatial deconvolution models using Smoother, namely non-negative least squares (NNLS), support vector regression (SVR)[19], dampened least squares (DWLS)[20], and log-normal regression (LNR)[21], to resolve the distribution pattern of cell types in a spatially informed manner (**Fig. 1a**). To benchmark the effect of Smoother, we first simulated ST data with distinct patterns, diverse degrees of spatial heterogeneity, and composition-independent structural noise (**Supplementary Fig. 5, 12 and Methods**). In almost every simulation scenario, the inclusion of spatial context leads to more accurate and realistic deconvolution results (**Fig. 1b-i, Supplementary Figs. 6-12**). This benefit is highly robust to the selection of models, hyperparameters, and marker genes, while the magnitude shrinks as cell-type-independent spatial dependencies grow (**Supplementary Figs. 13-15**). Surprisingly, we observed that CARD failed to take advantage of its own spatial modeling, and thus in some scenarios performed worse than a simple non-spatial model. Our results indicate that the parameterization of spatial structure used in CARD is suboptimal, echoing the importance of a unified and modular framework for dependency representation.

We next applied all spatially aware deconvolution methods to analyze real spatial omics data of normal, developmental and cancer tissues. Descriptions of each dataset and preprocessing details are provided in the Methods (**Methods**). We first evaluated deconvolution performance in detecting immune infiltration in a 10x Visium invasive ductal carcinoma dataset[22], where CD3 staining provided a ground truth for the presence of T cells. Smoother-guided NNLS accurately revealed the invasive lymphocytic pattern in tumor regions, whereas CARD failed to detect any T-cell signal inside the tumor (**Fig. 2a**). The benefit of spatial context was further validated by the elevated correlation of CD3 staining with estimated abundance and the improved binary prediction accuracy of T-cell presence (**Fig. 2b, c**). Second, to evaluate Smoother's ability to filter out noise while preserving meaningful boundaries, we performed deconvolution on a 10x Visium dataset of mouse brain[11] and estimated the abundances of 52 neural subtypes across brain locations simultaneously (**Fig. 2d**). As revealed by unsupervised clustering, Smoother notably reduced local discontinuities in cell-type composition, while the boundaries of tissue regions remained distinct (**Fig. 2e**). In particular, Smoother-based models clearly distinguished excitatory neuron subtypes in cortex and hippocampus, attenuating fluctuations within layers and subregions and producing sharper transitions compared to CARD (**Fig. 2f, g**).

Using a 50um spatial-CUT&Tag dataset of H3K4me3 in mouse E11 embryo[4], we further assessed the generalizability of Smoother-based deconvolution to epigenomics data (**Fig. 2h and Methods**). Smoother consistently produces more biologically coherent results as revealed by unsupervised clustering (**Fig. 2i**). In-depth examination of predicted cell-type composition and marker gene activity shows that the Smoother-guided approach has superior performance in restoring the spatial organization of the embryo with minimal levels of unsolved background (**Fig. 2j, k**). Specifically, the spatial structure of glial (*Sox9*), neural tube (*Ascl1*), cardiac muscle (*Hand2*), and definitive

erythroid (*Lmo2*) lineages are all clearly presented. Finally, we evaluated the scalability of Smoother-guided data imputation and deconvolution on datasets with higher spatial resolution. Using the Slide-seq2 data of human melanoma brain metastasis[23], we confirmed that Smoother, benefiting from the modular design and implementation, can enhance spatial patterns and provide accurate cell-type mappings to tens of thousands of locations while maintaining high computational efficiency (**Supplementary Fig. 16**).

Learning an informative low-dimensional representation is crucial for understanding the biological dynamics underlying noisy omics data. Smoother's ability to impose structural dependencies via a versatile spatial loss allows us to generalize existing non-spatial dimensionality reduction methods to spatial omics data (**Extended Data Fig. 2a, Methods and Supplementary Notes**). This is further strengthened by the contrastive extension of the spatial loss that better separates distant locations and prevents embeddings from collapsing. We first developed a variant of principal component analysis (PCA)[24] that leverages spatial context and applied it to encode the DLPFC data[10]. The quality of the low-dimensional representation is quantified by calculating the silhouette score using manual annotation as ground truth (**Extended Data Fig. 2b**). Across all 12 samples, the inclusion of spatial loss consistently improved model performance when the reconstruction and regularization were balanced, and the contrastive loss further increased the robustness by an order of magnitude (**Extended Data Fig. 2c, d**). These increments are directly visible in the 2D visualization where outer cortical layers are increasingly separated from inner layers and the white matter with increasing strength of regularization (**Extended Data Fig. 2e, f**). Similarly, we observed the same beneficial effect of spatial loss in a more general linear auto-encoder model where the symmetry and orthogonality constraints of PCA were relaxed (**Extended Data Fig. 2g-j**). Given the modularized structure of Smoother, we can extend any single-cell based non-spatial auto-encoder model to spatial data by fine-tuning over a new spatially aware objective. This provides an approach to transfer knowledge from single-cell atlases to spatial omics data.

In conclusion, we introduced Smoother to efficiently represent and exploit dependencies in spatial omics data. Smoother increases the computational burden only by a subtle margin (a few seconds for deconvoluting a typical 10x Visium ST dataset with ~5,000 spots), making it particularly well suited for emerging spatial omics techniques with better resolution and higher throughput. One caveat to Smoother is that the strength of imposed spatial dependencies relies upon prior assumptions on the data. Still, we demonstrated that Smoother is remarkably robust with respect to hyperparameters. It is also practicable to fine-tune the smoothing strength in specific tasks using cross validation and empirical Bayes approaches[25]. Collectively, Smoother is modular, scalable, and versatile to enhance a wide range of downstream tasks including data imputation, deconvolution, and dimensionality reduction. In light of the ubiquity of structural dependencies, we envision that Smoother may be readily extended to even more applications, such as trajectory inference and cell-cell communication, and may promote new biomedical discoveries in developmental and disease settings.

Online methods

Overview of Smoother

Smoother is a two-step framework that extracts prior dependency structures from positional information and integrates them into non-spatial models for downstream tasks to encourage local smoothness. We use $Y_{G\times S}$ to denote the spatial omics data of G genes (or the corresponding features in other modalities) at S locations, and $L_{F\times S}$ the spatial metadata of F spatial features (e.g., image feature extracted from histology) at S locations. The first step of Smoother is to construct a spatial graph G = (V, E, W) where the node set V represents locations and the edge set E and edge weights W are functions of $L_{F\times S}$. Then, Smoother derives a covariance structure $\Sigma_{S\times S}$ from the graph G and imposes it through the spatial prior on $Y_{G\times S}$ or other variables of interest. For the complete details of Smoother, see the **Supplementary Notes.**

Construction of spatial priors and losses.

Smoother draws from the concept of spatial stochastic process and a rich literature of spatial priors in image processing and geographic data analysis[9, 26, 27]. Intuitively, a spatial stochastic process allows us to model global dependency characteristics by assuming stationarity and specifying the local interactions of spatial random variables. Here we describe spatial connectivity using a weighted undirected graph G = (V, E, W) where physically adjacent locations (V) are connected by edges (E) with varying strengths (W). The spot-by-spot adjacency matrix $W_{S\times S}$, which is also called the spatial weights matrix of the underlying spatial process, specifies the connectivity (edge weights) between spots calculated from physical distance, where $w_{i,j} = 1$ if spot i and j are mutual k-nearest neighbors, otherwise 0. We set k=6 for data on hexagonal grids like 10x Visium. To encode domain boundaries, we further scale $W_{S\times S}$ using histological or transcriptomic pairwise similarities (soft-scaling), or manual domain annotations (hard-scaling):

$$W'_{S\times S} = W_{S\times S} \odot D_{S\times S}^{soft,hard}$$

where $D_{i,j}^{soft}$ is the pairwise similarity between spot i and j, and $D_{i,j}^{hard}$ is the binary indicator of whether spot i and j belong to the same domain. In practice, hard-scaling domains can be defined using segmentation of the histology image[28], clustering of transcriptomes or expert pathological annotations. For soft scaling, we extract per-spot gene expression or histological features and calculate $d_{i,j}$ as the pairwise similarity in a PCA-reduced space. If scaling the graph by transcriptomics, the first 10 PCs of gene expression and cosine similarity (zero for negative similarities) are used by default. If scaling by histology, the feature vector of a spot is the first three PCs of the concatenated RGB values of pixels in the square circumscribed about the spot, and similarity is calculated using Euclidean distance and a Gaussian kernel with band width 0.1 in the normalized PC score space.

Smoother next converts the spatial weights matrix $W_{S\times S}$ into a covariance structure $\Sigma_{S\times S}$ according to assumptions on the underlying stochastic process (see **Supplementary Notes**). We use the covariance to introduce a multivariate normal (MVN) prior on any spatial variable of interest, X_S. In particular, for a conditional autoregressive (CAR) process where the graph G describes a Gaussian Markov random field of X_S:

$$x_i \mid x_{\backslash i} \sim N\left(\rho\Sigma_{j\neq i}\, w_{ij}\, x_j, \sigma_i^2\right),$$

the joint distribution of X_S is a zero-centered MVN distribution with covariance $\Sigma_{S\times S}$, a smoothing prior that can be imposed on X_S in downstream tasks:

$$\Sigma_{S\times S} = diag(\sigma_i^{-2})(I - \rho W)^{-1}$$

$$P(X_S) \propto exp\left(-\frac{1}{2}\, X_S^T \Sigma_{S\times S}^{-1}\, X_S\right).$$

Here ρ is a scaling parameter to make $\Sigma_{S\times S}$ positive semi-definite and to control the decay rate of covariance over distance, which can be selected by examining the decay pattern of pairwise similarity. Since $\Sigma_{S\times S}$ is constructed from a boundary-aware graph, the above MVN prior provides a unique channel for neighboring locations to share information while still preserving boundaries.

To make Smoother compatible with existing non-spatial methods, we propose a quadratic regularization loss proportional to the density function of a zero-centered MVN with covariance $\Sigma_{S\times S}$:

$$L_{sp}(X_S; \Sigma_{S\times S}) = X_S^T \Sigma_{\mathrm{S\times S}}^{-1} X_S.$$

In general, for any given model with a loss function L_m, we may transform the model into a spatially aware version by minimizing a new joint loss function:

$$L_{joint}(X_S) = L_m(X_s) + \lambda_{sp} L_{sp}(X_S; \Sigma_{S\times S}).$$

The spatial loss term L_{sp} regularizes local fluctuations in X_s and makes the inference robust to technical noise. It can be shown that optimizing the new objective is equivalent to finding the maximum a posteriori (MAP) estimator with the spatial prior, and λ_{sp} can be viewed as the strength (inverse variance) of the prior. In addition, we implement a contrastive extension of the spatial loss to improve the separability of distant pairs of spots and prevent the inference from collapsing to trivial solutions. This is done by shuffling spot locations and generating corrupted covariance structures as negative samples,

$$L_{csp} \;=\; X_S^T\, \Sigma_0^{-1}\, X_S \;-\frac{1}{T}\sum_t X_S^T\, \Sigma_t^{-1}\, X_S$$

where Σ_0 is the covariance derived from the correct spatial graph and $\Sigma_t, t \in [1, T]$ are those from corrupted graphs.

Data imputation and resolution enhancement

Using local contexts and the prior dependency structure, Smoother is capable of de-noising and imputing spatial omics data at unseen locations. For a spatial random variable of interest, say X_S, we assume it follows a hidden Markov random field model and that technical noise is Gaussian iid., such that the observation Y_S is generated by the following model:

$$y_s \sim\; N(x_s, \sigma_0^2)$$
$$X_S \sim MVN(0, \sigma_1^2 \Sigma_{S\times S})$$

where σ_0^2 and σ_1^2 are the variance of observation and prior, respectively. Again, we can reparametrize the above problem and find the MAP estimator of X_S by solving the following optimization task with spatial loss:

$$\tilde{X}_{MAP} \;=\; argmin_X\, \|Y_S \;-\; X_S\|^2 \;+ \lambda_{sp}\, X_S^T\, \Sigma_{\mathrm{S\times S}}^{-1} X_S$$

where $\lambda_{sp} \propto \sigma_0^2/\sigma_1^2$ determines the strength of regularization. This is a special case of Tikhonov regularization and is equivalent to the weighted average filter commonly applied to smooth images. When part of the data is missing, the objective function is similar except the first reconstruction term is calculated only at observed positions. Since $\Sigma_{S\times S}$ is constructed a priori, Smoother can impute the latent value for an arbitrary set of locations and thus increase the spatial resolution. Note that the L2 reconstruction loss can be replaced by more complicated models to borrow information across features. See **Supplementary Notes** for further discussions.

Cell-type deconvolution

We define spatial deconvolution as the problem of inferring cell-type abundances at each location from the observed omics data, with or without cell-type reference information from external data. The task is especially relevant for spatial techniques with limited resolution, including 10x Visium, spatial-CUT&Tag[4] and spatial-ATAC-seq[5], where each profiled location represents a mixture of cells potentially from multiple cell types. The design of deconvolution model is usually determined by the generative model of the observations $Y_{G\times S}$. Most existing deconvolution methods assume a linear relationship between observation and cell-type abundance,

$$Y_{G\times S} \sim N(g(R_{G\times C}X_{C\times S}), \sigma_0^2)$$

where $Y_{G\times S}$ is the observed activities of G features at S spots, $R_{G\times C}$ is the expected reference activities of G features in C cell types, $X_{C\times S}$ is the abundance of C cell types at S spots, σ_0^2 is the sampling variability, and g is the data generative function that introduces additional noise, e.g., location-specific biases. Without loss of generality, we follow the same linearity assumption and extend non-spatial linear models to take advantage of positional information by imposing the spatial prior on the abundance of each cell type:

$$X_{c:} \sim MVN(0, \sigma_1^2\Sigma_c).$$

Here, Σ_c is the prior covariance structure, and $1/\sigma_1^2$ represents the strength of prior. In practice, when the reference $R_{G\times C}$ is known, usually from a paired single-cell dataset, we solve $X_{C\times S}$ by minimizing the following regularized factorization problem:

$$\tilde{X}_{MAP} = argmin_X\, L_{deconv}(Y\,, RX) + \lambda_{sp} \sum_c X_{c:}\Sigma^{-1}X_{c:}^T$$

where $L_{deconv}(Y\,, X\beta)$ is the loss specified by the corresponding deconvolution model. We implemented four spatially aware deconvolution models including non-negative least squares (NNLS), support vector regression (SVR), dampened least squares (DWLS), and log-normal regression (LNR). Detailed descriptions of these and more sophisticated models are provided in the **Supplementary Notes**. When the reference is unknown, the above deconvolution can be solved via matrix factorization, which is also a special case of the dimensionality reduction task, as will be discussed in the next section.

Dimensionality reduction

The goal of dimensionality reduction is to infer a low-dimensional representation that captures essential characteristics about the observed data while minimizing information loss. For spatial omics data, it is natural to believe that the latent dimensions should represent some continuous dynamics in space. Using Smoother, we can again impose prior and regularization on the hidden embedding to get more coherent representation of the data. We frame the dimensionality reduction task as finding a general auto-encoder model,

$$Y_s \xrightarrow{E_\theta} Z_s \xrightarrow{D_\phi} Y_s{}'$$

where E_θ is the decoder that projects omics data $Y_s \in R^G$ at the location s with G features onto a low-dimensional space R^H with H hidden dimensions, and D_ϕ is the decoder that projects the hidden embedding $Z_s \in R^H$ back to the original space. For any auto-encoder model with parameter θ, ϕ and reconstruction loss function $L_m(Y_s, \theta, \phi)$, Smoother incorporates spatial information by regularizing the hidden representation $Z_s(Y_s; \theta) = E_\theta(Y_s)$ using a spatial loss L_{sp} and solves a new joint objective

$$\hat{\theta}, \hat{\phi} = argmin_{\theta,\phi}\, L_m(Y, \theta, \phi) + \lambda_{sp}L_{sp}(E_\theta(Y); \Sigma_{S\times S}) + \lambda_c L_c(\theta, \phi)$$

where $L_c(\theta, \phi)$ is the additional soft constraint loss on model parameters, if any.

For linear dimensionality reduction models, including non-negative matrix factorization (NMF)[29], principal component analysis (PCA)[24], and independent component analysis (ICA)[30], the encoder E_θ and decoder D_ϕ are matrix multiplication operators and can be resolved via matrix factorization:

$$Y_{G\times S}' = W^{\phi}_{G\times H}\left(W^{\theta}_{H\times G} Y_{G\times S}\right)$$

$$\widetilde{W^{\theta}}, \widetilde{W^{\phi}} = argmin_{W^{\theta}, W^{\phi}}\ L_m\left(Y, W^{\phi}\left(W^{\theta}Y\right)\right) + \lambda_{sp}\, L_{sp}\left(W^{\theta}Y\right) + \lambda_c L_c\left(W^{\theta}, W^{\phi}\right).$$

The apparent analogy between the above linear auto-encoder (LAE) model and linear deconvolution model suggests that cell-type abundance itself can be viewed as the hidden factor. When the deconvolution reference $R_{G\times C}$ (i.e., $W^{\phi}_{G\times H}$ in the autoencoder) is unknown, factorizing $Y_{G\times S}$ is both a dimensionality reduction task and a reference-free deconvolution task. Further semi-supervised techniques can be applied to maximize the distinguishability of inferred latent states (cell types) using known marker genes.

In this study, we developed a variant of PCA and a relaxed LAE model in the Smoother framework to showcase the effect of spatial regularization. We assume all hidden dimensions to be independent and regularize them simultaneously. The PCA model has symmetric encoder and decoder, requires the latent dimensions to be orthogonal and uses L2 norm to measure the reconstruction loss. The relaxed LAE removes the symmetry and orthogonality requirements and includes an orthogonal loss to impose soft constraints on the latent embedding.

Model implementation

Smoother is publicly available as a Python package (https://github.com/JiayuSuPKU/Smoother/). Models involved in this study are implemented using PyTorch[31] and all optimizations are solved via PyTorch's gradient-based optimizers (by default Adam for deconvolution and SGD for dimensionality reduction). For convex problems, an alternative implementation of Smoother using CVXPY[32] is also available.

Recover spatial patterns of layer-specific genes in the human dorsolateral prefrontal cortex (DLPFC) dataset

We downloaded the preprocessed DLPFC data using the R package spatialLIBD[33] and used the 'logcounts' slot for data imputation. Specifically, we selected seven layer-specific marker genes reported in the original publication[10] (SNAP25, MOBP, PCP4, FABP7, PVALB, CCK, ENC1) and inferred the log-normalized expression level for each gene individually. The spatial prior was constructed under the intrinsic conditional autoregressive (ICAR) model with ρ = 0.99 and $\lambda_{sp} = 1$. For imputation, we randomly masked 1% to 20% of the total detected spots in a slide and used Smoother to estimate gene expression at masked locations. We quantified the performance of imputation by measuring the Pearson correlation between imputed data and real data. For resolution enhancement, we interpolated pseudo-spots uniformly across the surface with different resolutions and imputed expression for the new locations accordingly.

Extract spatial patterns in the metastatic melanoma Slide-seqV2 dataset

We preprocessed the Slide-seqV2 data of melanoma brain metastasis (MBM) and extracranial melanoma metastasis (ECM) and the paired scRNA-seq data following the original publication[23] using the R package Seurat[34]. Functional signature scores were calculated by averaging the

expression of genes involved in the corresponding pathway. We built an ICAR prior with $\rho = 0.9$ and imputed function scores individually using Smoother with $\lambda_{sp} = 1$ to mitigate dropout and to enhance the spatial resolution. Pairwise Pearson correlation between functional scores was calculated to evaluate the benefit of data imputation. We followed the original procedure to deconvolute the ST data using RCTD[13] and used the same set of RCTD-selected genes for Smoother-guided models.

Evaluate the benefit of Smoother on cell-type deconvolution using simulated and real spatial omics data

Simulate ST datasets with spatial patterns

We generate synthetic ST datasets following a procedure similar to that described in the cell2location paper[11], with key modifications enabling us to: (1) assign arbitrary spatial patterns to zones (co-localized cell-type groups) and (2) add additional noise to account for cell-type-independent dependencies. In brief, we first overlay designated patterns with a two-dimensional Gaussian process to generate the per-zone abundance values in space, then assign cell types to patterns and sample gene expression profiles at each location from a scRNA-seq reference according to the cell-type composition. The aggregated ST data is further combined with various sources of noise, including a lateral diffusion term to introduce confounding dependencies where each spot shares a certain proportion of mRNA to its neighbors directly, and a multiplicative per-gene gamma noise (by default shape = 0.4586, scale = 1/0.6992) to represent sampling error. The simulation code is provided as a stand-alone tool inside the Smoother package to facilitate future benchmarking efforts for ST related tasks.

Model comparison and evaluation

We focused our benchmarking on CARD since it is the only published spatially aware deconvolution method and has been shown to be superior to other existing methods[17]. However, Smoother is compatible with different deconvolution strategies. Any method, including the DWLS[20], nu-SVR[19], and LNR[21] re-implemented in this study, may be seamlessly transformed to take advantage of the spatial dependency structure. CARD contains a non-negative linear factorization model and introduces spatial dependencies by imposing a conditional autoregressive prior on cell-type abundances, which is mathematically similar to the Smoother-guided NNLS model. We followed the analysis tutorial and ran CARD with default parameters. In certain simulation scenarios, a truncated reference signature matrix with fewer genes is used as input directly to reflect limited shared information between external reference and the observed ST data. For the non-spatial CARD model, we set ρ to zero which effectively turned off any spatial interactions.

Deconvolution performance depends on the distribution properties of input data. For example, RNA-seq counts are typically skewed and, if not corrected, can bias the estimation against lowly expressed genes and rare cell types[21]. One simple solution is to perform deconvolution on the logarithmic scale, i.e., replacing Y and X with $log1p(Y)$ and $log1p(X)$. This approach is not physically sound, but empirically the log-scale has been shown to improve model performance by large margins, as do the square root scaling $\sqrt{Y}$ and $\sqrt{X}$ albeit to a less extent (data not shown). Unless otherwise stated, for benchmarking purposes we supplied log-scaled ST data to NNLS, DWLS, and nu-SVR and raw scale ST data to LNR and CARD. The reference expression matrix was calculated by taking the average of normalized expression of varying numbers of marker genes

across all cells of a given cell type from an external scRNA-seq dataset, followed by log-transformation where necessary. For more general cases, such as in epigenomics deconvolution where the reference and observation data may not be on the same scale or even from the same modality, we deliberately leave the choice of data scale to the user so that the model is best fit to the dataset.

For comparison purposes, in Smoother-guided models, we only used transcriptomic similarity to prune the spatial graph (KNN-based, k=6 for hexagonal grids and 4 for square grids) and applied the ICAR model (ρ = 0.99, which is the optimal ρ in CARD) to derive the covariance from the spatial weights matrix. The strength of the spatial prior λ_{sp} is set to 0 for all non-spatial baseline models, 1 for spatially aware NNLS, SVR, LNR, and 3 for spatially aware DWLS to adjust for the inflated model loss after scaling. This is not necessarily the best performing setting, as revealed by our parameter sensitivity analysis, but we stuck with the default since the benefit of spatial context is rather robust. Cell-type abundances were set to be non-negative in all models and were normalized to compute cell-type proportions at each spot as the final output.

Benchmark deconvolution performance using the simulated doughnut-shaped data

We obtained the single-nucleus RNA-seq data of mouse brain (5705STDY8058280, 5705STDY805828) along with cell type annotations for each cell from https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-11115/, and only kept the 15 most abundant neural subtypes for simulation. The dataset was further split into two, one to simulate ST data and the other as the deconvolution reference. We generated benchmarking ST datasets on a 50x50 grid containing five high-density (average of four cells per occupied spot) and 10 low-density (0.4 per occupied spot) cell types. Each cell type was assigned to a unique but overlapping doughnut-shaped distribution pattern. Cell-type-independent spatial dependencies were introduced by sharing 0/10%/50% of mRNA counts at each location to its first-degree neighbors. For each simulated dataset, we performed deconvolution using four sets of genes: (1) The union of the top 20 marker genes for each cell type (n=283), (2) the union of top 50 marker genes (n=656), (3) all informative markers genes whose log2(fold change) passes the threshold (by default 1) for one and only one cell type (n=2693), or (4) informative genes selected by the CARD model (n=6545). This in total brings 12 scenarios, each is replicated 10 times. All preprocessing and differential expression analysis steps were performed using the Python package Scanpy[35].

We evaluate deconvolution performance using three metrics: (1) The mean square error and (2) the Pearson correlation between ground truth and the estimated proportion per cell type, and (3) the average precision score of binary prediction of whether a cell type is present at a given location (abundance >= 0.1) using the Python package Scikit-learn[36]. Results across cell types and experiments (replicates) are aggregated together for visualization.

Benchmark deconvolution performance using the sci-Space data of mouse embryo

We obtained the sci-Space data[37] of 14 mouse embryonic slides from GEO under the accession number GSE166692. Each single cell in the sci-Space data is labeled with an approximate spatial coordinate (~200 um pitch). Cells of low-quality or from rare cell types (less than 100 cells in any slide) were filtered out. We averaged the expression level of 298 marker genes defined in the original paper to compute the deconvolution reference for the remaining eight cell types. To generate ST data of the same resolution, we pooled together all cells with the same coordinates

and performed resampling to adjust for the uneven density of captured cells. The actual cell count at each location is determined by a gamma-Poisson distribution with an expectation of 10. Since most coordinates have only one associated cell, we resampled cells with replacement from the current spot and adjacent locations (with half probability), effectively increasing the diversity at each spot. To make the deconvolution even harder, we down-sampled the synthetic ST data to have a maximum of 5,000 UMI per spot and added multiplicative per-gene gamma noise with mean around 0.35.

Benchmark deconvolution performance using the breast tumor dataset with staining

We obtained the invasive ductal carcinoma (IDC) data[22] with DAPI and anti-CD3 staining from 10x Genomics at https://support.10xgenomics.com/spatial-gene-expression/datasets and performed deconvolution to evaluate T-cell infiltration. The gene expression signature of nine cell types, including one tumor epithelial type, was computed using a scRNA-seq atlas of 26 breast cancer samples[38]. For Smoother-guided models, we selected the top 20 unique differentially expressed genes for each cell type using Seurat[34] and used them as input (180 genes in total) for deconvolution. For CARD, we provided the scRNA-seq data with all genes as input and run with default parameters. Performance was first evaluated by measuring the Spearman correlation between the CD3 immunofluorescence (IF) intensity and the T-cell deconvolution proportion over all spots on the slide. We then determined if these models could reliably assess T-cell infiltration in tumor regions (labeled as invasive carcinoma regions in the original publication), a question of paramount importance for immunotherapy research. Specifically, we fitted a two-component Gaussian mixture model on the CD3 intensity and predicted 377 out of the 2078 spots in the tumor regions to be T-cell positive. Binary prediction accuracy was measured by ROC score. In practice when the ground truth is not available, the prediction threshold of the abundance estimated by the deconvolution model can be established using the nuclei counts.

Visualize the smoothing effect using the 10x Visium mouse brain data

We obtained the 10x Visium dataset of mouse brain[11] and the paired snRNA-seq reference from ArrayExpress under accession IDs E-MTAB-11114 and E-MTAB-11115. We used the same cell-type annotation and removed seven unknown or low-quality cell types. The reference expression matrix for deconvolution was calculated for the remaining 52 neural subtypes. For Smoother-guided methods, the union set of the top 20 marker genes for each cell type discovered by Scanpy[35] was used to compute the reference matrix for deconvolution. For CARD, all QC-passed genes were supplied, and the reference matrix was estimated by CARD. Leiden clustering of estimated cell-type proportion was performed using Scanpy[35] to visualize brain regions with distinct cell-type compositions. Specifically, a KNN graph (k=38 to represent first- and second-degree neighbors in a hexagonal grid) was built using the first 20 PCs of the cell-type proportion. We manually adjusted the clustering resolution for different deconvolution methods so that the number of clusters generated was approximately the same. Clusters from different methods were aligned via linear sum assignment and colored accordingly.

Spatial mapping of scRNA-seq defined cell types in the spatial CUT&Tag data of mouse embryo

We obtained the spatial-CUT&Tag data and preprocessing scripts from the original publication[4]. Specifically, the R package ArchR[39] and the 'getGeneScore_ArchR.R' script were used to compute gene activity scores for the top 500 variable genes using the same set of parameters. We

removed spurious spots with outlier average gene activity (more than three standard deviations away from the mean), which lined up in one column or row on the grid and were considered as technical artifacts. The rest of the spots were used in all downstream analysis. For deconvolution, scRNA-seq data of E11 mouse embryo was obtained from Mouse Organogenesis Cell Atlas (MOCA)[40] and processed as described in the spatial-CUT&Tag paper. Specifically, we used the script 'integrative_data_analysis.R' to read and preprocess the subsampled (100k cells) data and Seurat's 'FindVariableFeatures' function to identify the top 500 variable genes. We then computed the pseudo-bulk expression count of variable genes for cell types annotated in 'Main_cell_type' group and used it as the deconvolution reference. Spatial domains of cell types were visualized using the same clustering strategy developed in the mouse brain analysis above.

Evaluate the benefit of Smoother on dimensionality reduction using the DLPFC dataset.
The log-normalized expression level of 1492 highly variable genes reported in the original DLPFC dataset[10] was used as input for dimensionality reduction. We constructed the spatial loss using the ICAR model and $\rho = 0.99$. All PCA and linear autoencoder models project the data onto 10 hidden dimensions and were initialized by SVD to speed up convergence. To construct the contrastive loss, we generated 20 corrupted graphs and set the relative importance of negative samples to 0.05. For linear autoencoder models, the strength of orthogonal loss λ_{orth} is set to 3. The silhouette score was computed using Scikit-learn[36], where distances were measured on the reduced 10-dimensional space.

Author contributions

J.S. conceived the original idea and developed the computational framework. R.R. and D.A.K. supervised the project. J.S., J.R., X.F., G.Z., and R.S.E carried out the analysis. J.J developed the data simulation scripts. Y.W. and B.I. contributed to the analysis of MBM data and the interpretation of the results. J.S. took the lead in writing the manuscript with input and critical feedback from all authors. All authors read and approved the final manuscript.

Acknowledgements

This work was funded by the National Institutes of Health, National Cancer Institute (R35CA253126 and U01CA243073 to R.R. and X.F., R37CA258829 and R21CA263381 to B.I.), the Center for Integrated Cellular Analysis (RM1HG011014 to J.S. and R.R.), and the Edward P. Evans Center for MDS at Columbia University (to J.S. and R.R.). The authors thank Y. Choi for contributions in benchmarking, P. Sims, Y. Shen, and C. Zhang for early discussions, and members of the Rabadan lab and Knowles lab for helpful comments and feedback throughout the project.

Declaration of Interests

R. Rabadan is the founder of Genotwin and consults for Arquimea Research. None of these activities are related to the results in the current manuscript.

Figure legends

Figure 1: Smoother: design and validation.

(a) Schematic overview of Smoother. Smoother first constructs a weighted spatial graph from physical positions, histology, and other features to represent spatial dependencies a priori (top). The prior is then imposed as a modular loss to regularize varying spatial variables, including gene activities, cell-type compositions, and latent embeddings (bottom).

(b-i) Smoother-guided deconvolution performance on simulated data. Doughnut-shaped ST datasets were generated from scRNA-seq reference[11] on a 50x50 grid where we assigned 15 cell types of high (n=5) and low (n=10) densities to overlapping spatial compartments **(b-e)**. ST datasets of mouse embryo were generated from the sci-Space dataset[37] by pooling single cells of eight types barcoded by one spot (~200um pitch) **(f-i)**. Relative abundance **(b, f)** and binary presence status (proportion >= 0.05, **d, h**) are visualized for selected cell types in one sample. Deconvolution accuracy is evaluated using Pearson correlation (**c, g**) and binary prediction average precision score (**e, i**). Error bars denote standard errors of the mean over 10 replicates for doughnut ST or 14 slides for sci-Space ST.

Figure 2: Smoother-guided cell-type deconvolution in various spatial omics data.

(a-c) Detection of tumor T-cell infiltration in a 10x Visium ductal carcinoma sample[22]. **(a)** Single channel images showing CD3 immunofluorescence staining (FITC/green) and estimated T-cell proportion (yellow) for Smoother-guided NNLS (left) and CARD (right). **(b)** Spearman correlation of predicted T-cell proportions with CD3 fluorescence levels across all spots on the slide. Whiskers indicate the 95% confidence intervals. **(c)** ROC score of the binary T-cell detection within tumor regions. Ground truth labels (visualized in **a**) were determined by a two-component Gaussian mixture model on CD3 intensity.

(d-g) Spatial mapping of neurons in the 10x Visium mouse brain slide ST8059048[11]. **(d)** Left: H&E staining of the brain section. Right: UMAP visualization of the paired single-nucleus transcriptomic reference of 52 neural subtypes. **(e)** Brain subregions identified by clustering over estimated cell-type compositions. **(f-g)** Estimated abundances of cortical **(f)** and hippocampal **(g)** excitatory neurons. Each color represents a regional neural subtype and its intensity is proportional to the estimated cell-type proportion, which is scaled by the same factor across methods for better visualization.

(**h-k**) Spatial mapping of cell types in the spatial-CUT&Tag H3K4me3 data of mouse embryo (50um)[4]. (**h**) Left: Overview of the technology adapted from[4]. Right: UMAP visualization of the scRNA-seq reference of 37 embryonic cell types[40]. (**i**) Embryonic subregions identified by clustering over estimated cell type compositions. (**j**) Estimated abundances of five major lineages aggregated from smaller cell types. Color intensities are scaled for better visualization. (**k**) Standardized gene activity of selected marker genes.

Extended Data Figure 1: Construction of spatial weights and covariance.

(**a**) Schematic overview. Smoother prunes the spatial graph using histology or other features to encode boundary information and to prevent unwanted cross-region interactions.

(b) H&E staining (left) and manual annotation of cortical layers (right) of slide 151673 of the DLPFC dataset[10].

(c) Covariance changes between neighboring pairs that belong to different layers (left) or the same layer (right) after complete (edge weight = 0, top) or leaky (edge weight = 0.01, bottom) hard scaling.

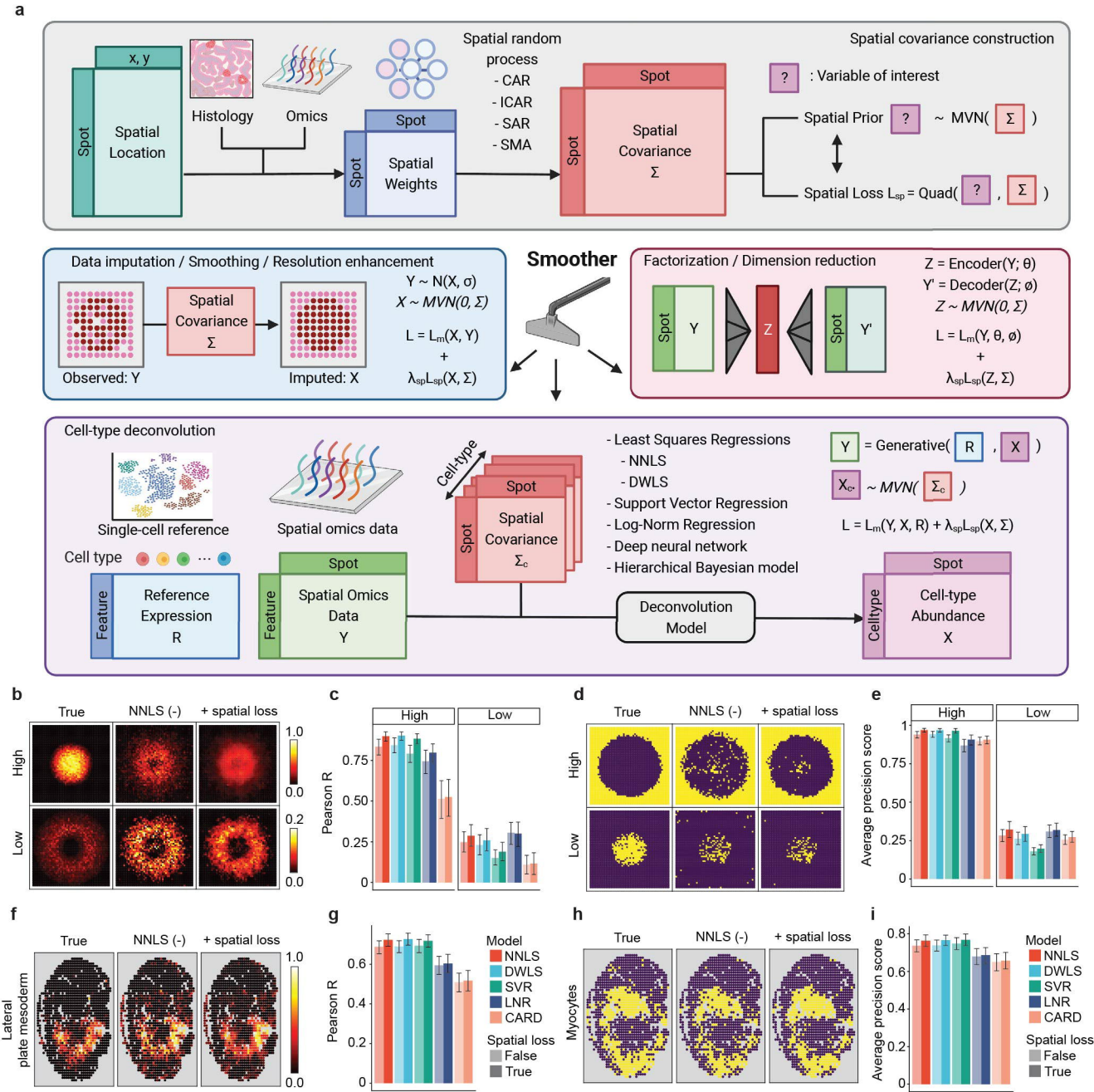
(d) Left to right: H&E staining, manual pathology annotation, histology clustering of H&E image and transcriptomics clustering of gene expression of cortical layers (right) of slide 1160920F of the human breast cancer dataset[41].
(e) Edge weight changes between neighboring pairs that belong to different pathological areas (left) or the same area (right) after scaling by different strategies. Top to bottom: soft-scaling by transcriptomic similarities, by histological similarities, hard-scaling by transcriptomics clusters, and by histological clusters.
(f) Covariance changes between neighboring pairs that belong to different pathological areas (left) or the same area (right) after scaling by different strategies, similar to **(e).**

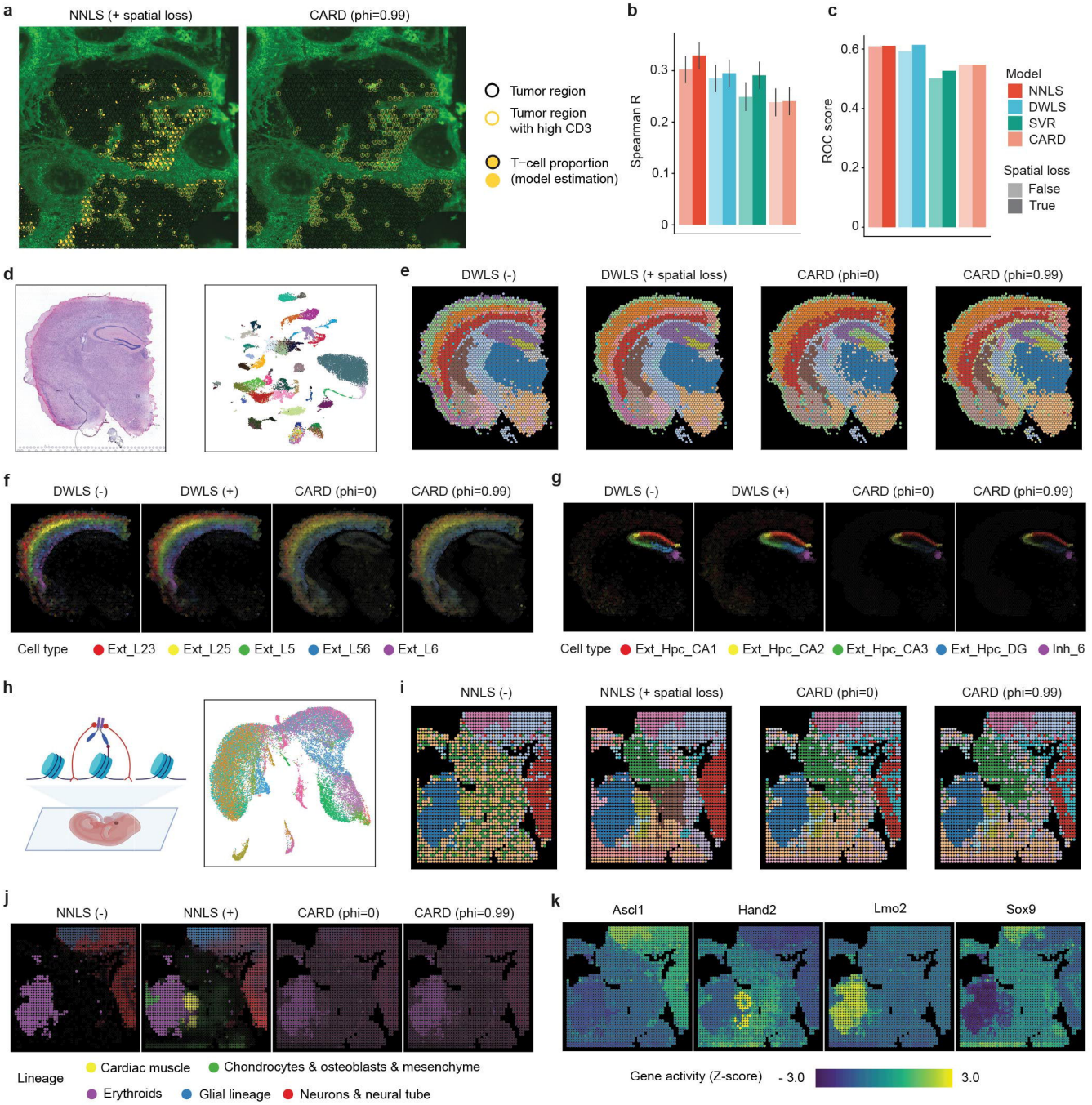
Extended Data Figure 2: Dimensionality reduction with spatial loss.
(a) Schematic overview of the spatially aware general auto-encoder framework. The overall objective consists of three parts: the reconstruction error, the spatial regularization loss, and any additional loss to impose constraints.
(b) Slide 151508 of the DLPFC dataset[10] colored by manual layer annotation.
(c-d) Dimensionality reduction performance of Smoother-guided PCA with varying strengths of regular **(c)** and contrastive **(d)** spatial losses. Silhouette scores were calculated based on the first 10 PCs and ground truth layer assignments. Performance gains were computed by subtracting the silhouette scores of non-spatial PCA results for each slide. Each boxplot shows the median and the $25^{th}/75^{th}$ percentiles as well as the above and below 1.5 times interquartile ranges indicated by the whiskers. The same below.
(e-f) Visualizations of the first two PCs of slide 151508 with regular **(e)** and contrastive **(f)** spatial losses. Color indicates layer annotation.
(g-h) Dimensionality reduction performance of a linear auto-encoder (LAE) as a relaxation of PCA with varying strengths of regular **(g)** and contrastive **(h)** spatial losses. The LAE is initialized using SVD to speed up convergence. Silhouette scores were calculated based on the 10 latent dimensions.
(i-j) Visualizations of the first two dimensions of slide 151508 with regular **(e)** and contrastive **(f)** spatial losses. Color indicates layer annotation.

References

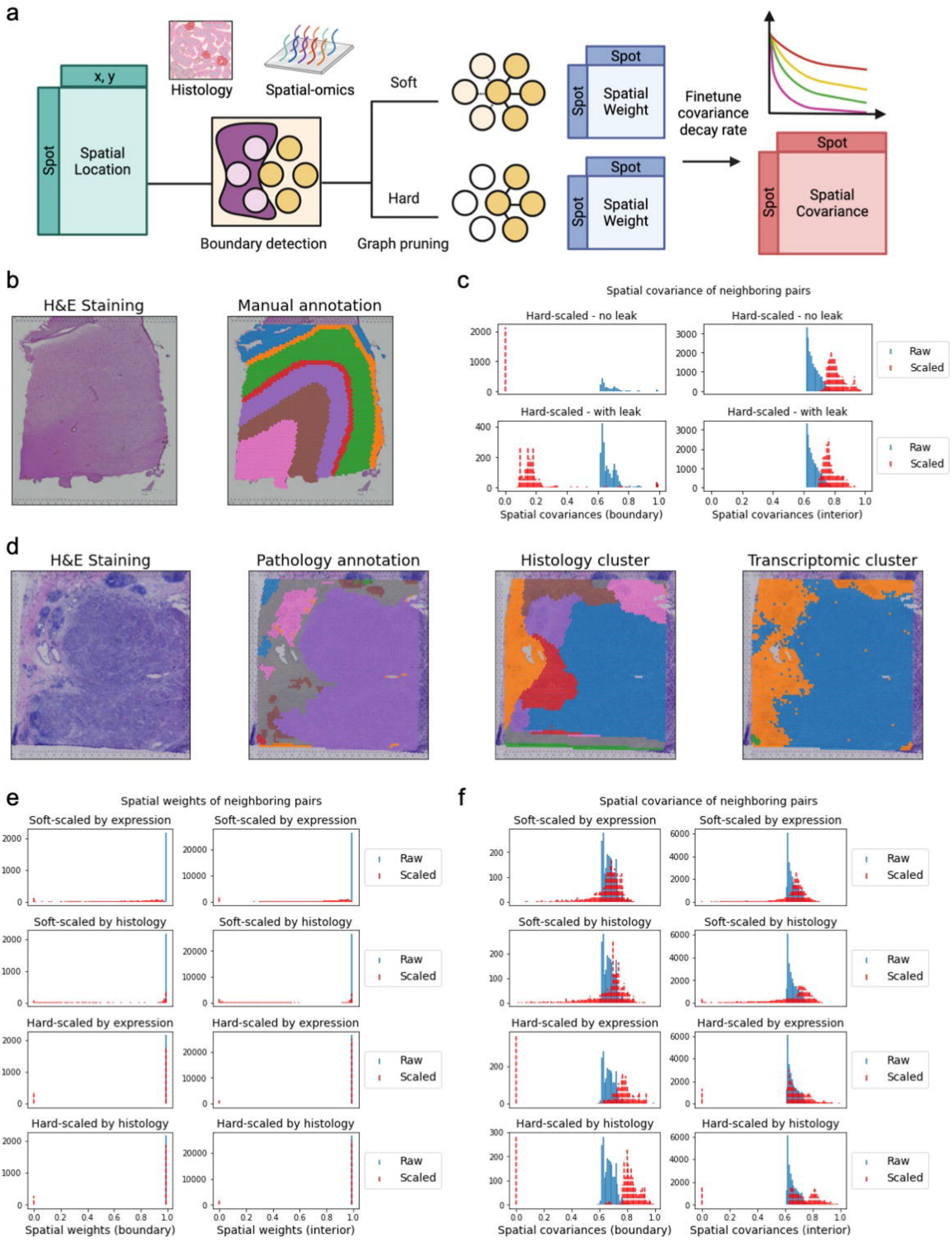
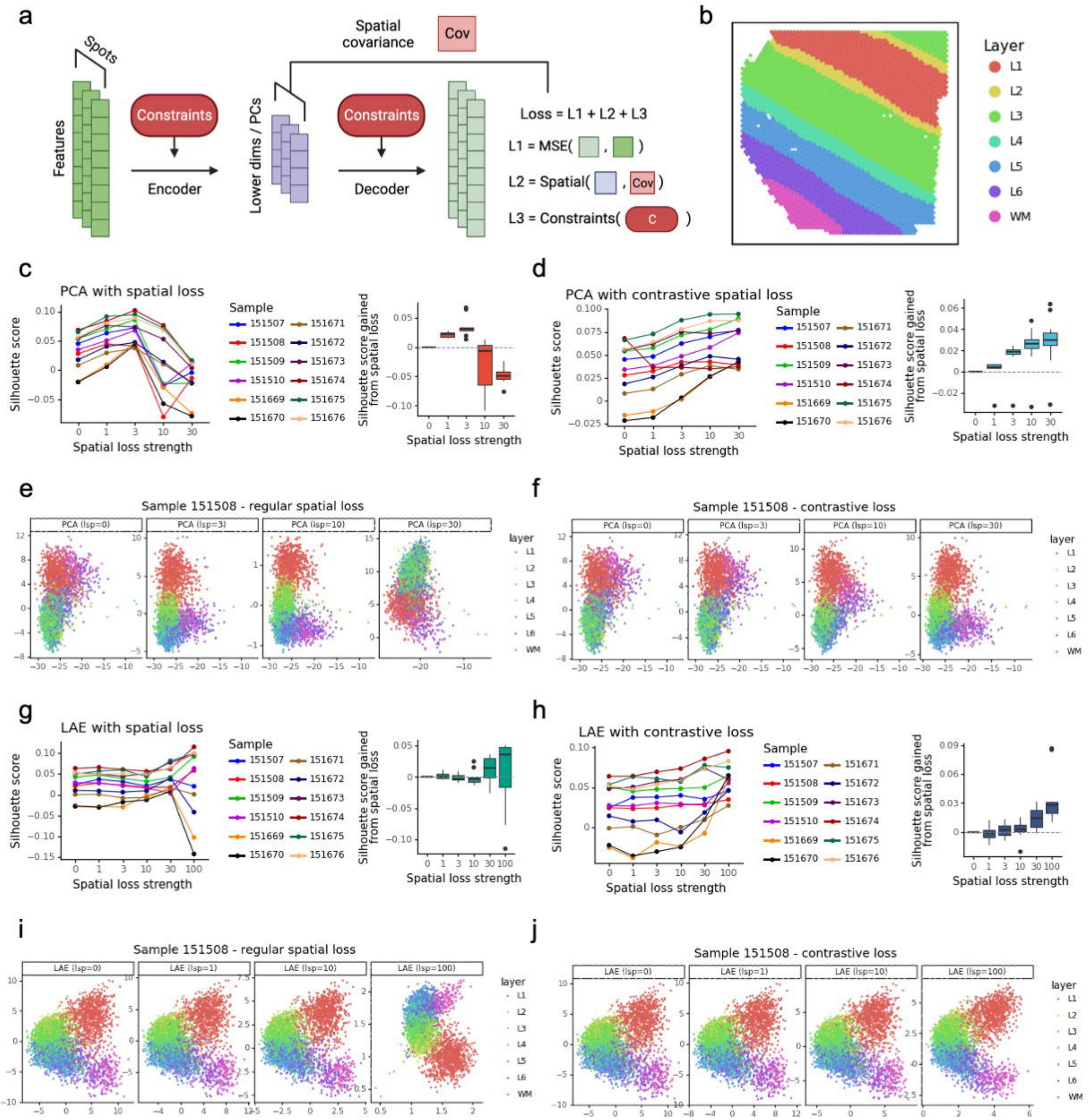
1. Moffitt, J.R., Lundberg, E. & Heyn, H. The emerging landscape of spatial profiling technologies. *Nature Reviews Genetics* (2022).
2. Stahl, P.L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78-82 (2016).
3. Stickels, R.R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat Biotechnol* **39**, 313-319 (2021).
4. Deng, Y.X. et al. Spatial-CUT&Tag: Spatially resolved chromatin modification profiling at the cellular level. *Science* **375**, 681-+ (2022).
5. Deng, Y.X. et al. Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* (2022).
6. Rao, A., Barkley, D., Franca, G.S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211-220 (2021).
7. Rue, H.v., Held, L. & ProQuest Gaussian Markov random fields : theory and applications. (Chapman & Hall/CRC, Boca Raton; 2005).
8. Sardy, S. & Tseng, P. On the statistical analysis of smoothing by maximizing dirty Markov random field posterior distributions. *J Am Stat Assoc* **99**, 191-204 (2004).
9. Besag, J., York, J. & Mollie, A. Bayesian Image-Restoration, with 2 Applications in Spatial Statistics. *Ann I Stat Math* **43**, 1-20 (1991).
10. Maynard, K.R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* **24**, 425-436 (2021).
11. Kleshchevnikov, V. et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* **40**, 661-+ (2022).
12. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res* **49** (2021).
13. Cable, D.M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat Biotechnol* **40**, 517-+ (2022).
14. Andersson, A. et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun Biol* **3**, 565 (2020).
15. Dong, R. & Yuan, G.C. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biol* **22**, 145 (2021).
16. Lopez, R. et al. DestVI identifies continuums of cell types in spatial transcriptomics data. *Nat Biotechnol* (2022).
17. Ma, Y. & Zhou, X. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat Biotechnol* **40**, 1349-+ (2022).
18. Zhang, H. et al. BayesTME: A unified statistical framework for spatial transcriptomics. *bioRxiv*, 2022.2007.2008.499377 (2022).
19. Newman, A.M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457 (2015).
20. Tsoucas, D. et al. Accurate estimation of cell-type composition from gene expression data. *Nat Commun* **10**, 2975 (2019).
21. Danaher, P. et al. Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data. *Nat Commun* **13** (2022).

22. Zhao, E. et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* **39**, 1375-+ (2021).
23. Biermann, J. et al. Dissecting the treatment-naive ecosystem of human melanoma brain metastasis. *Cell* **185**, 2591-+ (2022).
24. Plaut, E. From principal subspaces to principal components with linear autoencoders. *arXiv preprint arXiv:1804.10253* (2018).
25. Yang, L. & Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* **415**, 295-316 (2020).
26. Anselin, L. Spatial econometrics : methods and models, Edn. 1st 1988. (Kluwer Academic Publishers, Dordrecht, Netherlands; 1988).
27. van Erp, S., Oberski, D.L. & Mulder, J. Shrinkage priors for Bayesian penalized regression. *J Math Psychol* **89**, 31-50 (2019).
28. Palla, G. et al. Squidpy: a scalable framework for spatial omics analysis. *Nature Methods* **19**, 171-+ (2022).
29. Lee, D.D. & Seung, H.S. Algorithms for non-negative matrix factorization. *Adv Neur In* **13**, 556-562 (2001).
30. Comon, P. Independent Component Analysis, a New Concept. *Signal Process* **36**, 287-314 (1994).
31. Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems 32 (Nips 2019)* **32** (2019).
32. Diamond, S. & Boyd, S. CVXPY: A Python-Embedded Modeling Language for Convex Optimization. *J Mach Learn Res* **17** (2016).
33. Pardo, B. et al. spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. *Bmc Genomics* **23** (2022).
34. Hao, Y.H. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-+ (2021).
35. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19** (2018).
36. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825-2830 (2011).
37. Srivatsan, S.R. et al. Embryo-scale, single-cell spatial transcriptomics. *Science* **373**, 111-+ (2021).
38. Wu, S.Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* **53**, 1334-1347 (2021).
39. Granja, J.M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis (vol 53, pg 403, 2021). *Nat Genet* **53**, 935-935 (2021).
40. Cao, J.Y. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-+ (2019).
41. Wu, S.Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* **53**, 1334-+ (2021).


a
x, y
Spot
Spatial Location
Histology
Omics
Spot
Spot
Spatial Weights
Spatial random process
- CAR
- ICAR
- SAR
- SMA
Spot
Spot
Spatial Covariance Σ
Spatial covariance construction
? : Variable of interest
Spatial Prior ? ~ MVN(Σ)
Spatial Loss Lsp = Quad(? , Σ)
Data imputation / Smoothing / Resolution enhancement
Spatial Covariance Σ
Observed: Y
Imputed: X
Y ~ N(X, σ)
X ~ MVN(0, Σ)
L = Lm(X, Y) + λspLsp(X, Σ)
Smoother
Factorization / Dimension reduction
Spot
Y
Z
Spot
Y'
Z = Encoder(Y; θ)
Y' = Decoder(Z; ø)
Z ~ MVN(0, Σ)
L = Lm(Y, θ, ø) + λspLsp(Z, Σ)
Cell-type deconvolution
Single-cell reference
Cell type
Feature
Reference Expression R
Spatial omics data
Feature
Spot
Spatial Omics Data Y
Cell-type
Spot
Spot
Spatial Covariance Σc
- Least Squares Regressions
- NNLS
- DWLS
- Support Vector Regression
- Log-Norm Regression
- Deep neural network
- Hierarchical Bayesian model
Deconvolution Model
Y = Generative(R , X)
Xc ~ MVN(Σc)
L = Lm(Y, X, R) + λspLsp(X, Σ)
Spot
Celltype
Cell-type Abundance X
b
True
NNLS (-)
+ spatial loss
High
Low
1.0
0.0
0.2
0.0
c
High
Low
Pearson R
0.75
0.50
0.25
0
d
True
NNLS (-)
+ spatial loss
High
Low
e
High
Low
Average precision score
1
0.75
0.50
0.25
0
f
True
NNLS (-)
+ spatial loss
Lateral plate mesoderm
1.0
0.0
g
Pearson R
0.6
0.4
0.2
0
Model
NNLS
DWLS
SVR
LNR
CARD
Spatial loss
False
True
h
True
NNLS (-)
+ spatial loss
Myocytes
i
Average precision score
0.8
0.6
0.4
0.2
0
Model
NNLS
DWLS
SVR
LNR
CARD
Spatial loss
False
True


a
NNLS (+ spatial loss)
CARD (phi=0.99)
Tumor region
Tumor region with high CD3
T−cell proportion (model estimation)
b
Spearman R
0
0.1
0.2
0.3
c
ROC score
0
0.2
0.4
0.6
Model
NNLS
DWLS
SVR
CARD
Spatial loss
False
True
d
e
DWLS (-)
DWLS (+ spatial loss)
CARD (phi=0)
CARD (phi=0.99)
f
DWLS (-)
DWLS (+)
CARD (phi=0)
CARD (phi=0.99)
Cell type
Ext_L23
Ext_L25
Ext_L5
Ext_L56
Ext_L6
g
DWLS (-)
DWLS (+)
CARD (phi=0)
CARD (phi=0.99)
Cell type
Ext_Hpc_CA1
Ext_Hpc_CA2
Ext_Hpc_CA3
Ext_Hpc_DG
Inh_6
h
i
NNLS (-)
NNLS (+ spatial loss)
CARD (phi=0)
CARD (phi=0.99)
j
NNLS (-)
NNLS (+)
CARD (phi=0)
CARD (phi=0.99)
Lineage
Cardiac muscle
Chondrocytes & osteoblasts & mesenchyme
Erythroids
Glial lineage
Neurons & neural tube
k
Ascl1
Hand2
Lmo2
Sox9
Gene activity (Z-score)
- 3.0
3.0

Extended Data Fig. 1: Construction of spatial weights and covariance



Extended Data Fig. 2: Dimensionality reduction with spatial loss